Validation of remotely sensed profiles of atmospheric state variables: strategies and terminology

T. von Clarmann

Atmos. Chem. Phys. Discuss., 6, 4973–4994, 2006
www.atmos-chem-phys-discuss.net/6/4973/2006/

Atmospheric Chemistry and Physics Discussions

Validation of remotely sensed profiles of atmospheric state variables: strategies and terminology

T. von Clarmann

Forschungszentrum Karlsruhe, Institut für Meteorologie und Klimaforschung, Karlsruhe, Germany

Received: 18 April 2006 – Accepted: 2 May 2006 – Published: 20 June 2006

Correspondence to: T. von Clarmann (thomas.clarmann@imk.fzk.de)

Abstract

This technical note summarizes and classifies the various approaches to validation of remote measurements of atmospheric state variables, and tries to recommend a clear and unambiguous terminology. The following approaches have been identified: Intercomparison of single profiles for accuracy validation; statistical comparison of matched pairs of measurements with respect to bias determination and precision validation; statistical intercomparison of randomly sampled measurements by two instruments, and comparison of a single measurement to an ensemble of measurements. Applicable statistics are shortly reviewed, and recipes for evaluation of the co-incidence error due to less than perfect co-incidences are presented. A rigorous approach is suggested to quantitatively validate profile measurements when full covariance matrices are unavailable. We distinguish between "necessary validation" which is rejection of the null hypothesis that a difference between two measurements is significant, and "sufficient validation" which means to provide evidence that the probability that there is a significant difference is definitely small.

1 Introduction

Validation of a data product we understand is a statistical analysis of the differences between measurements of a new instrument to be validated, and a reference instrument already validated. The purpose is to detect any potential bias of the new measurement, and to verify that the estimated precision of the new measurement characterizes the measurements correctly.

Without any validated reference measurement available, it may also be helpful to intercompare measurements by two or more non-validated instruments. This approach we call "cross validation". While this approach certainly is no validation in its rigorous sense, it still may help to better characterize the data products.

ACPD

6, 4973–4994, 2006

2 Terminology

Let $\boldsymbol{x}=(x_1,\cdots,x_N)^T$ be a vertical profile of an atmospheric state variable, sampled on a discrete vertical grid of N altitude gridpoints, describing the true atmospheric state at the altitude resolution of the measurement to be validated. Let further $\hat{\boldsymbol{x}}=(\hat{x}_1,\cdots,\hat{x}_N)^T$ be a measurement of $\boldsymbol{x}$. The accuracy $\boldsymbol{a}$ of the measurement $\hat{\boldsymbol{x}}$ is the square root of the expectation value of the squared differences of the true quantities x_n and their measurements $\hat{x}_n$:

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} \sqrt{\langle(\hat{x}_1 - x_1)^2\rangle} \\ \sqrt{\langle(\hat{x}_2 - x_2)^2\rangle} \\ \vdots \\ \sqrt{\langle(\hat{x}_N - x_N)^2\rangle} \end{pmatrix} \tag{1}$$

The bias $\boldsymbol{b}$ of a measurement is the expection value of the deviation of the measured and the true quantity:

$$\boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{pmatrix} = \langle\hat{\boldsymbol{x}} - \boldsymbol{x}\rangle. \tag{2}$$

Depending on the nature of the bias, it can also be multiplicative rather than additive:

$$\boldsymbol{b}_{\text{mult.}} = \begin{pmatrix} b_{\text{mult.};1} \\ b_{\text{mult.};2} \\ \vdots \\ b_{\text{mult.};N} \end{pmatrix} = \left\langle \begin{pmatrix} \frac{\hat{x}_1}{x_1} - 1 \\ \frac{\hat{x}_2}{x_2} - 1 \\ \vdots \\ \frac{\hat{x}_N}{x_N} - 1 \end{pmatrix} \right\rangle \tag{3}$$

ACPD
6, 4973–4994, 2006

EGU

The precision $\boldsymbol{p}$ of the measurement characterizes the reproduceability of the measurement:

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{pmatrix} = \begin{pmatrix} \sqrt{\langle(\hat{x}_1 - \langle\hat{x}_1\rangle)^2\rangle} \\ \sqrt{\langle(\hat{x}_2 - \langle\hat{x}_2\rangle)^2\rangle} \\ \vdots \\ \sqrt{\langle(\hat{x}_N - \langle\hat{x}_N\rangle)^2\rangle} \end{pmatrix} \tag{4}$$

Accuracy, bias, and precision are related by

$$\begin{pmatrix} a_1^2 \\ a_2^2 \\ \vdots \\ a_N^2 \end{pmatrix} = \begin{pmatrix} b_1^2 \\ b_2^2 \\ \vdots \\ b_N^2 \end{pmatrix} + \begin{pmatrix} p_1^2 \\ p_2^2 \\ \vdots \\ p_N^2 \end{pmatrix} \tag{5}$$

Usually, remotely sensed data are provided along with careful data characterization, which includes estimates of the total error covariance matrix $\mathbf{S}_{\text{total}}$, the systematic error covariance matrix $\mathbf{S}_{\text{sys}}$ and the random error covariance matrix $\mathbf{S}_{\text{random}}$, and we have

$$\mathbf{S}_{\text{total}} = \mathbf{S}_{\text{sys}} + \mathbf{S}_{\text{random}}. \tag{6}$$

The diagonal elements of these matrices are the related variances $\sigma^2_{\text{total};n}=s_{\text{total};n,n}$, $\sigma^2_{\text{sys};n} = s_{\text{sys};n,n}$ and $\sigma^2_{\text{random}}=s_{\text{random};n,n}$, respectively. In the case of remote measurements, these error estimates typically are the linear mapping of known uncertainties (measurement noise, model parameter uncertainties etc) onto the retrieved quantities $\hat{x}_n$ (Rodgers, 2000, 1990). Validation then means to verify that for all n from 1 to N

$$a_n^2 = \sigma^2_{\text{total};n} \tag{7}$$

$$b_n^2 = \sigma^2_{\text{sys};n} \tag{8}$$

$$p_n^2 = \sigma^2_{\text{random};n}. \tag{9}$$

A useful strategy in validation is to first search for a possible bias, to suggest a bias correction, and to finally validate the estimated precision.

3 Comparison of co-incident measurements

3.1 General aspects

Let $\hat{\boldsymbol{x}}_{\text{val}}$ and $\hat{\boldsymbol{x}}_{\text{ref}}$ be two vertical profiles of the same quantity, measured by instruments "val" and "ref", respectiviely. The profiles and related diagnostic data have to be represented on a common grid, which usually implies regridding of one of both profiles (Calisesi et al., 2005). Further, if the measurements include a priori information, both profiles have to be transformed to the same a priori profile, and the smoothing error of the difference, $\mathbf{S}_{\text{smooth,diff}}$ has to be estimated (Rodgers and Connor, 2003). These authors suggest to quantify profile intercomparison by application of a χ^2 test:

$$\chi^2 = (\hat{\boldsymbol{x}}_{\text{val}} - \hat{\boldsymbol{x}}_{\text{ref}})^T \mathbf{S}_{\text{diff}}^{-1} (\hat{\boldsymbol{x}}_{\text{val}} - \hat{\boldsymbol{x}}_{\text{ref}}), \tag{10}$$

where $\mathbf{S}_{\text{diff}}$ is the covariance matrix of the difference with elements $s_{\text{diff};m,n}$ usually calculated as

$$\mathbf{S}_{\text{diff}} = \mathbf{S}_{\text{total,val}} + \mathbf{S}_{\text{total,ref}} + \mathbf{S}_{\text{coinc.}} + \mathbf{S}_{\text{smooth,diff}}, \tag{11}$$

unless both measurements have a common error source, which introduces correlations. This may apply to spectroscopic data uncertainties, similar temperature dependences of both measurement principles, etc. In case of such correlations, $\mathbf{S}_{\text{diff}}$ can be evaluated as

$$\mathbf{S}_{\text{diff}} = (\mathbf{I}, -\mathbf{I}) \begin{pmatrix} \mathbf{S}_{\text{total,val}}, \mathbf{C}_{\text{total,val,ref}} \\ \mathbf{C}^T_{\text{total,val,ref}}, \mathbf{S}_{\text{total,ref}} \end{pmatrix} (\mathbf{I}, -\mathbf{I})^T + \mathbf{S}_{\text{coinc.}} + \mathbf{S}_{\text{smooth,diff}}, \tag{12}$$

where $\mathbf{I}$ is $N \times N$ unity and where matrix $\mathbf{C}_{\text{total}}$ contains the related covariance elements between the new measurement "val" and the reference measurements "ref". In case

of less than perfect coincidences of the two measurements, it is important to quantify and consider the expectation value of the temporal and spatial coincidence error $\mathbf{S}_{\text{coinc.}}$ (see Sect. 3.2) (We avoid the term "mismatch error" in this context, because in parts of the community this term is reserved for the error which goes along with large values of the objective function of any kind of inverse modeling). Comparison of single profiles does not allow to distinguish between precision and bias validation.

3.2 Determination of coincidence error in time and space

Usually, only profiles are selected for comparison which meet a certain co-incidence criterion in time and space or any other adequate co-ordinates d like solar zenith angle, potential vorticity, equivalent latitude etc. The actual difference Δd in this quantity is the mismatch, and the maximum allowed mismatch is the co-incidence criterium Δd_{max}.

Variability of most atmospheric state variables is composed by a functional term and a random term. The abundance of a certain species, for example, may have a typical latitudinal dependence or a typical diurnal variation, which are superimposed by random fluctuations caused by the actual small-scale atmospheric situation. Whenever the mismatch is large enough for the functional dependence being important, and when the mismatch is large enough that nonlinear components of the functional dependence are important, the functional term should be corrected first by some appropriate parametrization $\boldsymbol{M}$. With d_{val} and d_{ref} being the co-ordinates of the validation and reference measurements, respectively, the uncorrected reference measurement $\hat{\boldsymbol{x}}_{\text{ref,uncorrected}}$ is corrected as

$$\hat{\boldsymbol{x}}_{\text{ref}} = \hat{\boldsymbol{x}}_{\text{ref,uncorrected}} + \boldsymbol{M}(d_{\text{val}}) - \boldsymbol{M}(d_{\text{ref}}). \quad (13)$$

and only the residual random term should be characterized by its covariance matrix. Otherwise the coincidence error may not follow a Gaussian distribution, and systematic sampling errors may inadvertently be treated as random coincidence errors. An example of application of a correction function $\boldsymbol{M}$ is found in Ridolfi et al. (2006)[1] who use

[1] Ridolfi, M., Blum, U., Carli, B., et al.: Geophysical Validation of temperature retrieved from

ECMWF temperature analyses to estimate the component of the differences between MIPAS and radiosonde temperatures which are explained by mismatch in space and time. A similar apprach was chosen by Cortesi et al. (2006)[2] for ozone.

To quantify the residual co-incidence error caused by finer structures than those accounted for by the correction function $\boldsymbol{M}$, a sufficiently fine resolved typical reference distribution $\hat{\boldsymbol{x}}_r$ of state variable $\boldsymbol{x}$ is needed. Let the reference distribution contain $l(\Delta d)$ independent pairs of data points separated by the mismatch $\Delta d = d_{val} - d_{ref}$. The coincidence error $\mathbf{S}_{coinc.}$ then can be evaluated as a function of Δd as

$$s_{coinc.;m,n}(\Delta d) = \frac{\sum_{i=1}^{l(\Delta d)} (\Delta \hat{x}_{r;m}(\Delta d))_i (\Delta \hat{x}_{r;n}(\Delta d))_i}{l - 1} - s_{distribution;m,n} \quad (14)$$

where

$$(\Delta \hat{x}_{r;m}(\Delta d))_i = (\hat{x}_{r;m}(d) - \hat{x}_{r;m}(d + \Delta d) - M_m(d) + M_m(d + \Delta d))_i \quad (15)$$

and

$$(\Delta \hat{x}_{r;n}(\Delta d))_i = (\hat{x}_{r;n}(d) - \hat{x}_{r;n}(d + \Delta d) - M_n(d) + M_n(d + \Delta d))_i \quad (16)$$

and where m and n identify the profile gridpoints, and where $s_{distribution;m,n}$ is an element of the random error covariance matrix $\mathbf{S}_{distribution}$ of the reference distribution of the state variable $\hat{x}_c$. $\mathbf{S}_{distribution}$ has to be estimated by error propagation calculation and cannot be obtained from the scatter of the reference sample, because the latter contains the natural variability we are trying to isolate. The M terms account for the difference already explained by the functional mismatch correction.

MIPAS/ENVISAT atmospheric Limb-emission measurements, Atmos. Chem. Phys. Discuss., in preparation, 2006.

[2]Cortesi, U., Blom, C., Blumenstock, Th., et al.: Co-ordinated validation activity and quality assessmentof MIPAS-ENVISAT Ozone data, Atmos. Chem. Phys. Discuss., in preparation, 2006.

In order to get $I(\Delta d)$ large enough for meaningful statistics, binning of $\mathbf{S}_{\text{coinc.}}$ is recommended, i.e. evaluation of $\mathbf{S}([\Delta d_1, \Delta d_2])$ for all mismatches in a range from Δd_1 to Δd_2, where $\mathbf{S}_{\text{coinc.}}$ is sufficiently linear in Δd. If such a bin $[\Delta d_1, \Delta d_2]$ covers the entire coincidence criterium, i.e. Δd_1=0 and Δd_2 equals the coincidence criterium, it is no longer necessary to care about the Δd-dependence of $\mathbf{S}_{\text{coinc.}}$ but the mean coincidence error $\overline{\mathbf{S}_{\text{coinc.}}} \approx \mathbf{S}_{\text{coinc.}}(\overline{\Delta d})$ can be used for the entire ensemble of coincidences.

Meteorological analyses, satellite measurements or modeled atmospheric fields can be used as reference distributions to evaluate the coincidence error on a larger scale. It is important to carefully assess any possible reduction of the horizontal variability in these datasets through application of Bayesian statistics in the sense of variational data assimilation (e.g. Ide et al., 1997) or optimal estimation retrievals (Rodgers, 1976). For determination of small-scale temporal fluctuations stationary in situ measurements or ground-based remote sensing measurements are better suited, while for small-scale spatial fluctuations aircraft measurements are the first choice.

Multi-dimensional co-incidence can be assessed component-wise by evaluation of Eq. (14) for each co-incidence direction (e.g. latitude, longitude and time) and summing up the respective coincidence error covariance matrices. In the case where the variation of the state variable under assessment is correlated between two of these dimensions, the summation has to be replaced by the following scheme:

$$\mathbf{S}_{\text{coinc.}} = (\mathbf{I}, \mathbf{I}) \begin{pmatrix} \mathbf{S}_{\text{coinc.};1}, \mathbf{C}_{\text{coinc.};1,2} \\ \mathbf{C}^{T}_{\text{coinc.};2,1}, \mathbf{S}_{\text{coinc.};2} \end{pmatrix} (\mathbf{I}, \mathbf{I})^{T}, \tag{17}$$

where the subscripts of the covariance matrices $\mathbf{S}_{\text{coinc.};l}$ and the cross-dimension covariances $\mathbf{C}_{\text{coinc.};k,l}$ denote the dimensions along which the variabilities are analyzed. Such correlations may apply, e.g., to the mixing ratio of an inert trace gas whose abundance is ruled by transport processes. The existence of a prevailing direction of wind in combination with a prevailing gradient in the field of the state variable then introduces such correlations.

Another option to handle co-incidence errors in L dimensions is to define

a norm of the following type which transforms the multi-dimensional mismatch $\Delta\boldsymbol{d}=(\Delta d_1, \ldots, \Delta d_L)$ to a scalar mismatch distance Δd:

$$\Delta d = \sqrt{\sum_{1}^{L} (w_l \Delta d_l)^2}, \tag{18}$$

where w_l are weighting factors reflecting the expected variability of the state variable with the respective direction l. Steck et al. (2006)[3], e.g., have used

$$\Delta d = \sqrt{\Delta_{\text{long}}^2 + \Delta_{\text{lat}}^2 + (\Delta_t v_w)^2} \tag{19}$$

where Δ_{long} and Δ_{lat} are longitudinal and latitudinal mismatch distances, Δ_t is the mismatch in time, v_w is the typical windspeed. This particular norm holds for analysis of transport-dominated abundances of trace species without prevailing gradients and wind directions.

3.3 Horizontal smoothing

Additional complication arises if the measurements to be compared characterize air parcels of non-zero extension in the direction of d. In this case the smoothing error in direction of d and the coincidence error can no longer be treated independent. First we discuss the along-line-of-sight extension of an air parcel sounded by a limb-viewing instrument. If the extension of an air parcel sounded by the measurement system to

[3]Steck, T., Blumenstock, T., Clarmann, T., Glatthor, N., Grabowski, U., Hase, F., Hochschild, G., Höpfner, M., Kellmann, S., Kiefer, M., Kopp, G., Linden, A., Milz, M., Oelhaf, H., Stiller, G. P., Wetzel, G., Zhang, G., Fischer, H., Funke, B., Wand, D. Y., Gathen, P., Hansen, G., Stebel, K., Kyrö, E., Allaart, M., Redondas Marrero, A., Remsberg, E., Russell III, J., Steinbrecht, W., Yela, M., and Raffalski, U.: Validation of ozone measurements from MIPAS-Envisat, J. Geophys. Res., under review, 2006.

be validated is larger in the direction of d than that represented by an element of the reference measurement $\boldsymbol{x}_r$, then the profiles $\boldsymbol{x}_r$ in Eq. (14) have to be replaced by

$$\tilde{\boldsymbol{x}}_r = \mathbf{A}_{hor}\boldsymbol{x}_r \tag{20}$$

Further, the covariance matrix $\mathbf{S}_{distribution}$ has to be replaced by

$$\tilde{\mathbf{S}}_{distribution} = \mathbf{A}_{hor}\mathbf{S}_{distribution}\mathbf{A}_{hor}^T, \tag{21}$$

where $\mathbf{A}_{hor}$ is the matrix of horizontal averaging kernels in the line-of-sight direction. These can be obtained e.g. from perturbationan analysis or analytically from 2D radiative transfer modelling and retrieval tools (see, e.g. Steck et al., 2005; Carlotti et al., 2001). If $\mathbf{A}_{hor}$ is not available, it can be approximated by $\mathbf{RA}$, where $\mathbf{A}$ is the vertical profile averaging kernel matrix, and $\mathbf{R}$ the $I \times N$ dimensional ray-tracing operator, which maps altitudes $z_1, \cdots, z_n$ to along-track distances $d_1, \cdots, d_I$ according to the observation geometry. Elements of $\mathbf{A}$ representing contributions from below the tangent altitude are assigned to the tangent point geolocation. This approximation, however, neglects both the mapping of any horizontal smoothing error onto the retrieved profile, and the asymmetry of the horizontal averaging kernel around the tangent point of a limb viewing measurement. This approach has been chosen by Ridolfi et al. (2006)[1] and Cortesi et al. (2006)[2] to account for the horizontal smoothing of MIPAS in the co-incidence correction.

To account for the cross-line-of-sight extension of the air-parcel sounded by a limb viewing instrument, $\mathbf{R}$ is unity and $\mathbf{A}$ is the horizontal cross-line-of-sight field-of-view weighting function. For nadir sounding instruments, the latter approach can be applied in either direction.

For comparison of a limb viewing measurement with a measurement of negligible horizontal smoothing, the pure relative smoothing error without any co-incidence error component can be evaluated by application of the approach proposed above in this Section (Eqs. 20 and 21) to Eq. (14) for Δd=0. Similar considerations apply to comparison of two limb sounders with different azimuth viewing direction. In this case

the smoothing errors of both instruments have to be combined under consideration of correlations if the lines of sight are not orthogonal.

4 Bias determination

To determine the bias between two measurement systems, a statistical ensemble of measurements is needed. This ensemble can either be composed of K matching pairs of measurements or random samples of K and L measuremets of each measurement system, respectively.

4.1 Statistical bias determination with matching pairs of measurements

The mean difference between measurements to be validated and coincident reference measurements can be compared with its statistical uncertainty in order to determine any bias between the measurement to be validated and the reference measurement and its significance. With K pairs of coincident measurements available, the bias $\boldsymbol{b}$ is estimated at

$$\hat{\boldsymbol{b}} = \frac{\sum_{k=1}^{K}(\hat{\boldsymbol{x}}_{\mathrm{val};k} - \hat{\boldsymbol{x}}_{\mathrm{ref};k})}{K} \tag{22}$$

The statistical uncertainty of the bias is characterized by the related covariance matrix $\mathbf{S}_{\mathrm{bias}}$, of which the elements are estimated at

$$s_{\mathrm{bias};m,n} = \frac{\sum_{k=1}^{K}(\hat{x}_{\mathrm{val};m,k} - \overline{\hat{x}}_{\mathrm{val};m})(\hat{x}_{\mathrm{val};n,k} - \overline{\hat{x}}_{\mathrm{val};n})}{K(K-1)} + \frac{\sum_{k=1}^{K}(\hat{x}_{\mathrm{ref};m,k} - \overline{\hat{x}}_{\mathrm{ref};m})(\hat{x}_{\mathrm{ref};n,k} - \overline{\hat{x}}_{\mathrm{ref};n})}{K(K-1)} - \frac{\sum_{k=1}^{K}(\hat{x}_{\mathrm{val};m,k} - \overline{\hat{x}}_{\mathrm{val};m})(\hat{x}_{\mathrm{ref};n,k} - \overline{\hat{x}}_{\mathrm{ref};n})}{K(K-1)} - \tag{23}$$

$$= \frac{\sum_{k=1}^{K}(\hat{x}_{\mathrm{val};m,k} - \hat{x}_{\mathrm{ref};m,k} - \hat{b}_m)(\hat{x}_{\mathrm{val};n,k} - \hat{x}_{\mathrm{ref};n,k} - \hat{b}_n)}{K(K-1)},$$

where

$$\bar{\hat{x}} = \frac{\sum_{k=1}^{K}\hat{x}_k}{K}. \tag{24}$$

This assessment does not need any error estimates of $\hat{\boldsymbol{x}}_{\mathrm{val}}$ or $\hat{\boldsymbol{x}}_{\mathrm{ref}}$. If, however, precision estimates of differences

$$\mathbf{S}_{\mathrm{diff,random}} = \mathbf{S}_{\mathrm{val,random}} + \mathbf{S}_{\mathrm{ref,random}} + \mathbf{S}_{\mathrm{val,coinc.}} \tag{25}$$

are available, the measurements can be weighted accordingly in the bias determination:

$$\hat{\boldsymbol{b}} = \left(\sum_{k=1}^{K}\mathbf{S}^{-1}_{\mathrm{diff,random};k}\right)^{-1}\left(\sum_{k=1}^{K}\mathbf{S}^{-1}_{\mathrm{diff,random};k}(\hat{\boldsymbol{x}}_{\mathrm{val};k} - \hat{\boldsymbol{x}}_{\mathrm{ref};k})\right) \tag{26}$$

The bias uncertainty in terms of covariance matrix then is

$$\mathbf{S}_{\mathrm{bias}} = \left(\sum_{k=1}^{K}\mathbf{S}^{-1}_{\mathrm{diff,random};k}\right)^{-1} \tag{27}$$

The consistence of $\hat{b}_n$ and $\sigma_{\mathrm{sys},n}$ can easily be checked (see, e.g. Ridolfi et al., 2006[1], for application to MIPAS temperature validation, or Cortesi et al., 2006[2], for ozone validation). Evaluation of the significance of the bias then requires χ^2 statistics, where

$$\chi^2_{\mathrm{bias}} = \hat{\boldsymbol{b}}^T\mathbf{S}_{\mathrm{bias}}\hat{\boldsymbol{b}}. \tag{28}$$

Obviously, neither the root mean squares difference of profiles obtained from two measurement systems nor $1/\sqrt{K}$ of the root mean squares difference are a measure of the significance of the bias.

The extension of the bias determination discussed above to the determination of a multiplicative bias is straightforward: The mean relative deviation of a state parameter x_n at altitude gridpoint n is calculated as

$$\hat{b}_{\text{mult.};n} = \frac{\sum_{k=1}^{K} \frac{\hat{x}_{\text{val};n,k} - \hat{x}_{\text{ref};n,k}}{\hat{x}_{\text{ref};n,k}}}{K}. \tag{29}$$

The elements of its covariance matrix are calculated as

$$s_{\text{bias,mult.};m,n} = \frac{\sum_{k=1}^{K} \left(\frac{\hat{x}_{\text{val};m,k} - \hat{x}_{\text{ref};m,k}}{\hat{x}_{\text{ref};m,k}} - \hat{b}_{\text{mult.};m}\right)\left(\frac{\hat{x}_{\text{val};n,k} - \hat{x}_{\text{ref};n,k}}{\hat{x}_{\text{ref};n,k}} - \hat{b}_{\text{mult.};n}\right)}{K(K-1)}, \tag{30}$$

The multiplicative bias is the natural choice to report the systematic errors which are expected to be proportional to the state parameter itself. Retrieval errors due to erroneous line intensities in spectrometric remote sensing of trace gas abundances are a typical example. Some authors prefer to report relative mean deviation instead:

$$\hat{b}_{\text{rel.};n} = \frac{\hat{b}_n}{\frac{\sum_{k=1}^{K} \hat{x}_{\text{ref},k}}{K}} = \frac{\hat{b}_n}{\bar{\hat{x}}_{\text{ref},n}} \tag{31}$$

This way to report the bias is adequate when the nature of the bias is additive and the ratioing by the reference value (used here as an estimate of the true value) only serves the purpose to illustrate the relevance of this error component with respect to a typical value. The estimation of the variance $s_{b,\text{rel};n,n}$ of this quantity requires error propagation calculation under consideration of correlations, because the counter and the denominator include common terms:

$$\sigma^2_{b,\text{rel};n,n} = s_{b,\text{rel};n,n} = \frac{1}{\bar{\hat{x}}^2_{\text{ref};n}}\left(s_{\text{bias};n,n}\bar{\hat{x}}^2_{\text{ref};n} + s_{\text{ref};n,n}\hat{b}^2_n - 2r\sigma_{\text{bias};n,n}\sigma_{\text{ref};n,n}\bar{\hat{x}}_{\text{ref},n}\hat{b}_n\right), \tag{32}$$

where $\sigma_{\text{bias};n,n} = \sqrt{s_{\text{bias};n,n}}$ is the standard deviation of the bias; $\sigma_{\text{ref};n,n}$ is the standard deviation of the mean reference value given by

$$\sigma^2_{\text{ref};n,n} = s_{\text{ref};n,n} = \frac{\sum_{k=1}^{K}(\hat{x}_{\text{ref},n} - \bar{\hat{x}}_{\text{ref},n})^2}{K(K-1)} \tag{33}$$

and r is the correlation coefficient of counter and denominator,

$$r = \frac{\sum_{k=1}^{K}(\hat{x}_{\text{val};k,n} - \hat{x}_{\text{ref};k,n} - \hat{b}_{k,n})(\hat{x}_{\text{ref};k,n} - \bar{\hat{x}}_{\text{ref},n})}{(K-1)\sigma_{\text{bias};n,n}\sigma_{\text{ref};n,n}}. \tag{34}$$

The simplified expression

$$\tilde{\sigma}_{b,\text{rel};n,n} = \frac{\sigma_{\text{bias};n,n}}{\bar{\hat{x}}_{\text{ref},n}} \tag{35}$$

ignores the uncertainty of $\bar{\hat{x}}_{\text{ref},n}$.

4.2 Bias determination by statistical comparison of random samples

It is not necessary to use matched pairs for validation. Random samples are sufficient but any sampling artefacts have to be carefully excluded. A parametrization as suggested in Sect. 3.2, Eq. (13) may help to reduce systematic sampling errors.

When two instruments provide large but independent, i.e. unmatched, random samples of measurements, the bias can be determined as the difference of respective mean values:

$$\hat{\boldsymbol{b}} = \frac{\sum_{k=1}^{K}\hat{\boldsymbol{x}}_{\text{val};k}}{K} - \frac{\sum_{l=1}^{L}\hat{\boldsymbol{x}}_{\text{ref};k}}{L} = \bar{\hat{\boldsymbol{x}}}_{\text{val}} - \bar{\hat{\boldsymbol{x}}}_{\text{ref}} \tag{36}$$

and the respective covariance matrix has the elements

$$s_{\text{bias};m,n} = \frac{\sum_{k=1}^{K}(\hat{x}_{\text{val};m} - \bar{\hat{x}}_{\text{val};m})(\hat{x}_{\text{val},n} - \bar{\hat{x}}_{\text{val};n})}{K(K-1)} + \tag{37}$$

$$\frac{\sum_{l=1}^{L}(\hat{x}_{\mathrm{ref},m} - \overline{\hat{x}}_{\mathrm{ref};m})(\hat{x}_{\mathrm{ref},n} - \overline{\hat{x}}_{\mathrm{ref};n})}{L(L-1)}.$$

Obviously, any non-randomness of the samples can cause an apparent bias.

5 Precision validation

5.1 Precision determination with matching pairs of measurements

For accuracy validation, the root mean squares difference of the pairs of matched measurements is calculated, which we expect to equal the total estimated error of the profile difference. In terms of variances and covariances, we test that

$$\langle \sum_{k=1}^{K}(\hat{x}_{\mathrm{val};k,m} - \hat{x}_{\mathrm{ref};k,m})(\hat{x}_{\mathrm{val};k,n} - \hat{x}_{\mathrm{ref};k,n})\rangle = s_{\mathrm{diff};m,n}, \quad (38)$$

which again is to be verified by χ^2 statistics:

$$\langle(\hat{\boldsymbol{x}}_{\mathrm{val}} - \hat{\boldsymbol{x}}_{\mathrm{ref}})^T \mathbf{S}_{\mathrm{diff}}^{-1}(\hat{\boldsymbol{x}}_{\mathrm{val}} - \hat{\boldsymbol{x}}_{\mathrm{ref}})\rangle = N \quad (39)$$

However, if there is a bias $\boldsymbol{b}$ between the measurement systems, this should be removed in order to validate the precision of the measurement rather than the accuracy. This leads to the following χ^2 test

$$\langle(\hat{\boldsymbol{x}}_{\mathrm{val}} - \hat{\boldsymbol{x}}_{\mathrm{ref}} - \hat{\boldsymbol{b}})^T \mathbf{S}_{\mathrm{diff,random}}^{-1}(\hat{\boldsymbol{x}}_{\mathrm{val}} - \hat{\boldsymbol{x}}_{\mathrm{ref}} - \hat{\boldsymbol{b}})\rangle = N - 1, \quad (40)$$

where $\mathbf{S}_{\mathrm{diff,random}}$ is the random component of $\mathbf{S}_{\mathrm{diff}}$ according to Eq. (12) (see, e.g. Ridolfi et al., 2006[1], for application to MIPAS temperature validation, or Cortesi et al., 2006[2], for ozone validation).

The variances and covariances of the bias-corrected differences are related to the variances and covariances of the bias by

$$\sum_{k=1}^{K}(\hat{x}_{\mathrm{val};k,m}-\hat{x}_{\mathrm{ref};k,m}-\hat{b}_m)(\hat{x}_{\mathrm{val};k,n}-\hat{x}_{\mathrm{ref};k,n}-\hat{b}_n)=K s^2_{\mathrm{bias};m,n}, \tag{41}$$

where K is the sample size.

5.2 Precision validation by comparison of random samples

The scatter of a sample of measurements is composed of both the measurement random error (characterized by covariance matrices $\mathbf{S}_{\mathrm{random;val}}$ or $\mathbf{S}_{\mathrm{random;ref}}$, respectively) and the natural variability (characterized by its covariance matrix $\mathbf{S}_{\mathrm{nat}}$). The natural variability of two randomly sampled data sets, however, is the same, regardless if we observe the atmosphere with the one or the other instrument. Thus, we have to verify

$$\mathbf{S}_{\mathrm{val;nat}}=\mathbf{S}_{\mathrm{val;sample}}-\mathbf{S}_{\mathrm{val;random}}=\mathbf{S}_{\mathrm{ref;sample}}-\mathbf{S}_{\mathrm{ref;random}}=\mathbf{S}_{\mathrm{ref;nat}}, \tag{42}$$

where the elements of $\mathbf{S}_{\mathrm{val;sample}}$ are

$$s_{m,n}=\frac{\sum_{k=1}^{K}(\hat{x}_{\mathrm{val};m}-\overline{\hat{x}}_{\mathrm{val};m})(\hat{x}_{\mathrm{nal};n}-\overline{\hat{x}}_{\mathrm{val};n})}{K-1} \tag{43}$$

and where the elements of $\mathbf{S}_{\mathrm{ref;sample}}$ are

$$s_{m,n}=\frac{\sum_{l=1}^{L}(\hat{x}_{\mathrm{ref};m}-\overline{\hat{x}}_{\mathrm{ref};m})(\hat{x}_{\mathrm{ref};n}-\overline{\hat{x}}_{\mathrm{ref};n})}{L-1} \tag{44}$$

Testing of $\mathbf{S}_{\mathrm{val;nat}}$=$\mathbf{S}_{\mathrm{ref;nat}}$ is performed with the F-test (see, e.g., Press et al., 1989). The strategy discussed here is particular sensitive to an artificial reduction of the variability of one of the measurement data sets through the use of retrieval schemes involving Bayesian statistics, where each single profile is pushed towards some a priori information.

6 Comparison of a single measurement with a random sample of measurements

If only single profile measurements are available which do not coincide with any of the measurements to be validated, it can be checked if the single profile measurement could be part of the distribution defined sample of size K to be validated. The applicable χ^2 test then is

$$(\overline{\hat{\boldsymbol{x}}_{\text{val}}} - \hat{\boldsymbol{x}}_{\text{ref}})^T (\mathbf{S}_{\text{val;ensemble}} + \mathbf{S}_{\text{ref;total}})(\overline{\hat{\boldsymbol{x}}_{\text{val}}} - \hat{\boldsymbol{x}}_{\text{ref}}) \tag{45}$$

where $\mathbf{S}_{\text{val;ensemble}}$ is the ensemble covariance matrix of the measurements to be validated. Its elements are

$$s_{\text{val;ensemble};m,n} = \sum_{k=1}^{K} \frac{(\hat{x}_{\text{val};m,k} - \overline{\hat{x}_{\text{val};m}})(\hat{x}_{\text{val};n,k} - \overline{\hat{x}_{\text{val};n}})}{K-1}. \tag{46}$$

Again, considerations as outlined above Eq. (12) may apply.

7 How much validation is enough?

The rationale of validation often is as follows: The null hypothesis is that the difference of the profiles is significant. χ^2 statistics allows to calculate the probability $P_{\text{acc}}(\chi^2)$ that the actual χ^2 occurs accidently, due to the error bars, without any substantial difference of the profiles. If this probability is larger than a given threshold, corresponding to an actual χ^2 below a given threshold, the null hypothesis of a significant difference has to be rejected. The integral of the χ^2 probability density function from the actual χ^2 to infinity may result in a value of slightly above, say, 0.05, which implies a probability $P_{\text{dis}}(\chi^2)$ of a substantial, i.e., non-accidental disagreement below 95%. The null hypothesis of substantial disagreement is thus not significant at 5% confidence level and must be rejected. However, we only know that the probability $P_{\text{acc}}(\chi^2)$ that there

is agreement between the measurements is larger than 5%, which does not suggest that the new data are really trustworthy. The failure of proving significance of disagreement is not equivalent with the evidence of agreement. We call this level of validation "necessary validation" but it is indeed only failed falsification.

If, however, the (predefined!) intercomparison ensemble contains K comparisons, with all χ^2_k smaller than a critical threshold χ^2_{crit} representing the 5-% confidence level, than the probability of disagreement $P_{\text{dis}}(K, \chi^2_{\text{max}})$ is for independent profile measurements according to the multiplication axiom

$$1 - P_{\text{acc}}(K, \chi^2_{\text{max}}) = P_{\text{dis}}(K, \chi^2_{\text{max}}) < 0.95^K, \quad (47)$$

where χ^2_{max} is the largest χ^2 value found in the ensemble. With an ensemble of enough intercomparisons (K=59 for the 5-%-threshold), each with a $\chi^2_k < \chi^2_{\text{crit}}$ equivalent to a probability of disagreement $P_{\text{dis}}(k, \chi^2_k)$ of less than 95%, the probability $P_{\text{acc}}(\chi^2_{\text{max}}, K)$ that there is no significant disagreement is larger than 95%, and the new measurement is validated at 5% confidence level. We call this level of validation "sufficient validation". With a large maximum χ^2_{max} in the comparison ensemble, a large ensemble is needed with all χ^2_k below the threshold χ^2_{max}, while with a smaller maximum χ^2_{max}, a smaller ensemble is sufficient. Similar considerations apply to the F-test using for statistical validation of random rather than matched samples.

Alternatively one can also perform a single χ^2 test for the entity of measurements. For horizontally uncorrelated measurements the total χ^2 simply is the sum of the individual profile χ^2 values. The expectation value of χ^2 then is the degree of freedom of the whole system, i.e. $N \times K$. In this case the probability $P_{\text{dis}}(\chi^2)$ of substantial disagreement of the entire comparison ensemble can be evaluated directly by integration of the χ^2 probability density function.

The advantages of the approach involving the multiplication axiom are (1) for the lower degree of freedom the user will more easily find tabulated values of P in the literature; (2) the availability of numerous χ^2 values allows to verify that these follow the expected distribution. The major disadvantage is that this probability estimate is

not “sufficient” (in estimation theoretical terminology; beware of the different meaning compared to “sufficient condition”!), because it is based only on the maximal χ^2_{max} of the entire ensemble but not on the actual values of the other ones. Further, it is not “robust” because it is very sensitive to outliers. The advantages of the χ^2 test of the entire comparison ensemble is its inherent sufficiency. Disadvantages are on the practical side because the integrals of the χ^2 probability density functions are not usually tabulated for large degrees of freedom. The safest is to combine both approaches. Discrepancies can then point at non-representative outliers in the comparison ensemble. E.g., Migliorini et al. (2004) have detected suspicious ozone profiles in their comparison ensemble by comparison of the expected and the found χ^2 distribution.

If the null hypothesis cannot be rejected (failed necessary validation), or if no sufficient confidence in the agreement of the new measurement and the reference measurement can be achieved (failed sufficient validation), the precision estimates of one of the instruments have been too optimistic, or the coincidence error may have been underestimated, or any other excuse will be found by the responsible scientist in charge.

In the case the sufficient validation fails, there are three options: 1. One option is not to change anything and just to report a poorer level of significance. This is option of choice if no further validation measurements are available and if the largest χ^2_{max} values cannot be explained by outliers which are not part of the expected χ^2 distribution and which cannot be sorted out with good reason, as discussed under option (3). 2. If the initial ensemble size was chosen too small, the profile causing χ^2_{max} may not be representative. In this case, a larger comparison ensemble may help to achieve a reasonable significance level according to Eq. (47). If, however, the initial sample was representative, even larger χ^2 values will occur in the larger ensemble, and the significance level will not improve. It is, of course, important to work with pre-defined random samples and not to adjust the sample or the sample size to the optimum significance level. 3. Large χ^2 can also be associated with a particular subset of the sample which can be characterized by some objective criterion. Migliorini et al. (2004), e.g. have found problems in O_3 data from spectra suspected to be cloud contaminated. In this

ACPD
6, 4973–4994, 2006

EGU

case it may be appropriate to define a kind of data filter and to validate only the complementary subset which passes the filter. There are, however, two traps in this approach: First, the filter should not use the quantity to be validated itself as a filter criterium. Second, the new analysis system, of which the newly defined filter is a part, has to be validated using an independent comparison ensemble. When the original sample is used, it will always be possible to tune the data filter such that a good significance level is achieved.

8 What if full retrieval error covariance matrices are not available?

Without the profile covariances available, we cannot draw any quantitative conclusion on the reliability of the retrieved profiles in the sense of χ^2 statistics. Often, however, after debiasing, there are at least no horizontal error correlations to be considered. Then, state variables can be compared and χ^2 statistics can be set up for a large ensemble of size K of scalar measurements to be validated $\hat{x}_{\mathrm{val},n,k}$ and reference measurements $\hat{x}_{\mathrm{ref},n,k}$ at a single selected altitude $z(n)$. This corresponds to "map validation" instead of "profile validation". All formulation discussed in this paper then is applied to the simple case where N=1. χ^2-testing in the sense as discussed in Section 7 in this application leads to a valid conclusion on the reliability of a measurement $\hat{x}_{\mathrm{val},n}$ at a certain confidence level 1-P at the selected altitude $z(n)$. Of course, this procedure can be performed for all altitudes of interest independently. We consider a profile measurement system sufficiently validated if we can sufficiently validate the values at each altitude. If, after debiasing, correlations in the time domain can be excluded, the rationale outlined above also can be applied to time series validation. Ridolfi et al. (2006)[1] have combined the map validation and time series validation approach by statistically analyzing differences between MIPAS temperatures and radiosonde temperatures from two stations measured at various times. The statistical analysis was peformed for altitude bins defined such that each MIPAS limb scan (i.e. each profile) was represented only once in each bin, justifying to disregard any error correlations in altitude.

9 Conclusions

Recipes and terminology for statistical validation of a profile measurement system have been suggested which cover both bias and precision validation and which are applicable to both matched pairs of coincident measurements and random samples of measurements. Further, a recipe has been suggested to validate profile measurements in a statistical rigorous way even if their full profile covariance matrices are not available. The rejection of the hypothesis that there is a set of reference measurements which is significantly in contradiction with the measurements to be validated is a necessary condition of validation. The more rigorous approach to give evidence that the probability of the existence of a significant contradictory measurement is small is considered a sufficient condition. While in real life it will not always be possible to apply these approaches at full rigorosity, validation scientists certainly will find workarounds and simplifications. It is hoped that this technical note at least supports better communication in the validation community by suggesting a more or less consistent terminology. Further, ad hoc validation approaches may serve their purpose better, once clarified which rigorous approach they are meant to replace.

Acknowledgements. The author would like to thank all people who patiently listened to his thoughts even at times of the day when statistics are not usually discussed.

References

Calisesi, Y., Soebijanta, V. T., and van Oss, R.: Regridding of remote soundings: Formulation and application to ozone profile comparison, J. Geophys. Res., 110, D23306, doi:10.1029/2005JD006122, 2005. 4977

Carlotti, M., Dinelli, B. M., Raspollini, P., and Ridolfi, M.: Geo-fit approach to the analysis of limb-scanning satellite measurements, Appl. Opt., 40, 1872–1885, 2001. 4982

Ide, K., Courtier, P., Ghil, M., and Lorenc, A. C.: Unified Notation for Data Assimilation: Operational, Sequential and Variational, J. Meteorolog. Soc. Japan, 75, 1B, 1997. 4980

Migliorini, S., Piccolo, C., and Rodgers, C. D.: Intercomparison of direct and indirect measurements: Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) versus sonde ozone profiles, J. Geophys. Res., 109, D19316, doi:10.1029/2004JD004988, 2004. 4991

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T.: Numerical Recipes, Cambridge University Press, Cambridge, 1989. 4988

Rodgers, C. D.: Retrieval of Atmospheric Temperature and Composition From Remote Measurements of Thermal Radiation, Rev. Geophys. Space Phys., 14, 609–624, 1976. 4980

Rodgers, C. D.: Characterization and error analysis of profiles retrieved from remote sounding measurements, J. Geophys. Res., 95, 5587–5595, 1990. 4976

Rodgers, C. D.: Inverse Methods for Atmospheric Sounding: Theory and Practice, vol. 2 of Series on Atmospheric, Oceanic and Planetary Physics, edited by: Taylor, F. W., World Scientific, 2000. 4976

Rodgers, C. D. and Connor, B. J.: Intercomparison of remote sounding instruments, J. Geophys. Res., 108, 4116, doi:10.1029/2002JD002299, 2003. 4977

Steck, T., Höpfner, M., von Clarmann, T., and Grabowski, U.: Tomographic retrieval of atmospheric parameters from infrared limb emission observations, Appl. Opt., 44, 3291–3301, 2005. 4982